

# The Reactome pathway knowledgebase

David Croft<sup>1</sup>, Antonio Fabregat Mundo<sup>1</sup>, Robin Haw<sup>2</sup>, Marija Milacic<sup>2</sup>, Joel Weiser<sup>2</sup>, Guanming Wu<sup>2</sup>, Michael Caudy<sup>2</sup>, Phani Garapati<sup>1</sup>, Marc Gillespie<sup>3</sup>, Maulik R. Kamdar<sup>2</sup>, Bijay Jassal<sup>2</sup>, Steven Jupe<sup>1</sup>, Lisa Matthews<sup>4</sup>, Bruce May<sup>2</sup>, Stanislav Palatnik<sup>2</sup>, Karen Rothfels<sup>2</sup>, Veronica Shamovsky<sup>4</sup>, Heeyeon Song<sup>2</sup>, Mark Williams<sup>1</sup>, Ewan Birney<sup>1</sup>, Henning Hermjakob<sup>1,\*</sup>, Lincoln Stein<sup>2,5,6,\*</sup> and Peter D'Eustachio<sup>4,\*</sup>

<sup>1</sup>European Bioinformatics Institute (EMBL-EBI), European Molecular Biology Laboratory, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK, <sup>2</sup>Ontario Institute for Cancer Research, Toronto, ON M5G0A3, Canada, <sup>3</sup>College of Pharmacy and Health Sciences, St. John's University, Queens, NY 11439, USA, <sup>4</sup>NYU School of Medicine, New York, NY 10016, USA, <sup>5</sup>Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA and <sup>6</sup>Department of Molecular Genetics, University of Toronto, Toronto, ON M5S 1A1, Canada

Received October 1, 2013; Accepted October 19, 2013

## ABSTRACT

**Reactome (<http://www.reactome.org>) is a manually curated open-source open-data resource of human pathways and reactions. The current version 46 describes 7088 human proteins (34% of the predicted human proteome), participating in 6744 reactions based on data extracted from 15 107 research publications with PubMed links. The Reactome Web site and analysis tool set have been completely redesigned to increase speed, flexibility and user friendliness. The data model has been extended to support annotation of disease processes due to infectious agents and to mutation.**

## INTRODUCTION

At the cellular level, life is a network of molecular reactions that can be organized into higher order interconnected pathways. Molecules are synthesized, degraded, transported from one location to another and assembled into complexes and higher order structures with other molecules. Intensive studies of cellular signaling, motility, vesicular trafficking and other aspects of cell biology, coupled with the development of comprehensive catalogs of human genes and their protein products, have enabled the description of many cellular processes in the same molecular detail that has been a standard for metabolic processes for a generation. By annotating all of these processes in a single, consistent reaction-pathway format, the Reactome Knowledgebase systematically links human

proteins to their molecular functions, providing a resource that functions both as an archive of biological processes and as a tool for discovering unexpected functional relationships in data from gene expression pattern surveys or somatic mutation catalogues from tumor cells (e.g. 1–3).

Since its inception 10 years ago, Reactome has grown to include (version 46–September 2013) annotations for 7088 of the 20 774 protein-coding genes in the current Ensembl human genome assembly (34% coverage), 15 107 literature references and 1421 small molecules organized into 6744 reactions collected in 1481 pathways. Notable recent additions include extensive annotations of phospholipid and eicosanoid metabolism, protein glycosylation and SCF-KIT, IGF1R, NOTCH and HIPPO signaling, as well as annotations of regulatory processes mediated by non-coding RNAs.

Here, we will focus on three new features of Reactome: the development of a strategy to annotate the disease counterparts of normal human processes, the deployment of a redesigned Web site and the extension of tools for data analysis.

### Disease curation using an enhanced data model

The Reactome data model (4) builds on earlier work by Kanehisa *et al.* (5) and Karp *et al.* (6) to classify and catalog physical entities (proteins and other macromolecules, small molecules, complexes of these entities and post-translationally modified forms of them), their subcellular locations and the transformations they can undergo (biochemical reaction, association to form a complex and translocation from one cellular compartment

\*To whom correspondence should be addressed. Tel: +1 212 263 5779; Fax: +1 212 263 8166; Email: Peter.D'Eustachio@nyumc.org  
Correspondence may also be addressed to Henning Hermjakob. Tel: +44 1223 494 671; Fax: +44 1223 494 468; Email: hhe@ebi.ac.uk  
Correspondence may also be addressed to Lincoln Stein. Tel: +1 416 673 8514; Email: Lincoln.Stein@oicr.on.ca

The authors wish it to be known that, in their opinion, the first six authors should be regarded as Joint First Authors.

to another). The central class of the Reactome data model is a *Reaction*, and subclasses of *Reaction* model these core biological events. Reactions are grouped into pathways, which in turn are assembled into a hierarchy of biological processes. Wherever appropriate, Reactome entities are linked to external reference databases such as UniProt (7), Ensembl (8), ChEBI (9) and Rhea (10). The full Reactome database schema is available at [http://www.reactome.org/cgi-bin/classbrowser?DB=gk\\_current](http://www.reactome.org/cgi-bin/classbrowser?DB=gk_current).

A straightforward definition of disease allows us to annotate a broad range of major disease processes at the molecular level. Diseases that can now be annotated in Reactome arise in one of three ways: a mutation, somatic or germ-line, leads to a non-functional gene product so processes that normally depend on that gene product do not take place; a mutation leads to a gene product with a novel function, enabling novel reactions whose products perturb normal human processes; or an infectious agent such as a virus introduces novel gene products whose novel reactions perturb normal human processes.

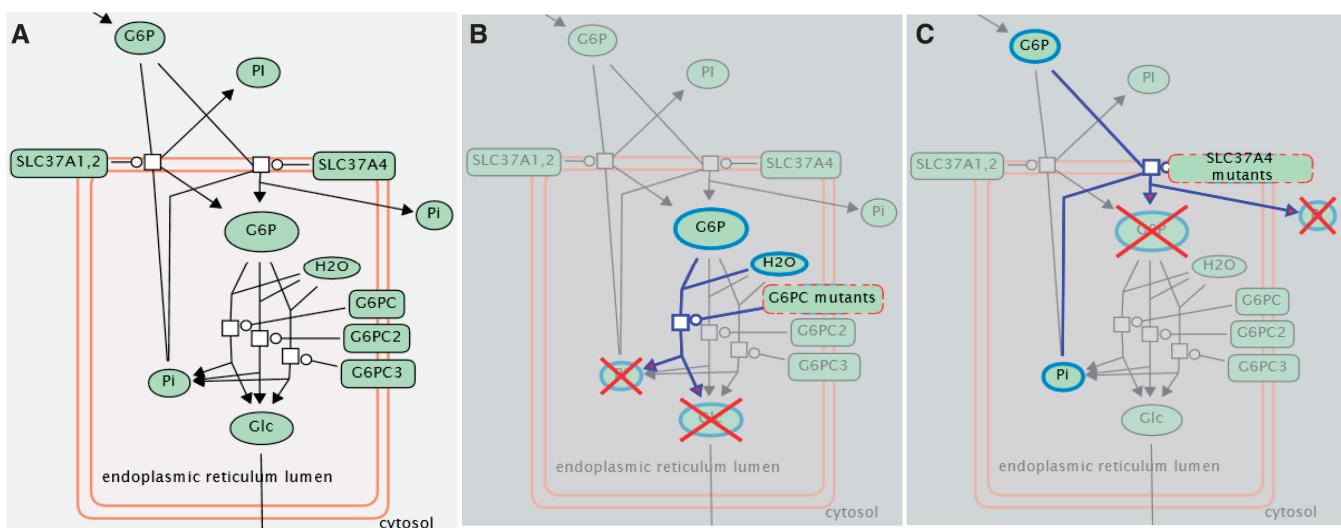
To identify disease-associated entities and events, a new ‘disease’ attribute is added, taking its value terms from a disease ontology [currently <http://disease-ontology.org/> (11)]. This attribute is multivalued, so an entity or event that has roles in multiple disease processes can be annotated to capture all of those roles.

With this addition, diseases due to infection can be annotated within our data structure. Our existing ‘species’ attribute allows pathogen-derived proteins, DNA and RNA to be distinguished from molecules encoded in the human genome. The ‘species’ attribute, also associated with complexes, reactions and pathways, can be multivalued, allowing complexes containing both host and viral proteins or reactions involving host and viral components to be properly identified. Finally, addition of Gene Ontology (GO) (12) host\_cell terms to our cell compartment vocabulary allows us to localize pathogen-derived entities accurately, in compliance with GO annotation practice.

To denote molecular properties of a protein modified by a somatic or germline mutation in the gene that encodes it, as opposed to post translational modification, we created a new class. This class of genetic modifications has subclasses to accommodate substitution of a canonical residue by a different one, the insertion or deletion of multiple contiguous residues into a canonical sequence and the generation of a fusion protein containing fragments of two canonical ones (13). The underlying chemical similarity between a protein that differs from its canonical form due to a mutation and one that differs due to co- or post-translational modification allows us to maintain compliance with the PSI-MOD standard for annotation of protein modifications (14).

Reactions involving mutated proteins as catalysts, inputs, outputs and regulators are annotated exactly as wild-type reactions are. To link reactions involving a mutated protein to those involving its normal counterpart, an optional ‘normal reaction’ attribute is used. A reaction in which a receptor constitutively activated by a mutation transmits a signal is thereby paired with the wild-type reaction in which a normal receptor is activated by ligand binding. A reaction catalyzed by the normal form of an enzyme is paired with the different one catalyzed by its gain-of-function mutant counterpart or with a dead-end reaction (normal inputs, no outputs) associated with its loss-of-function mutant counterpart. A limitation of this strategy at present is that, as Reactome does not capture quantitative data such as reaction rates or binding affinities, quantitative effects of mutations are not readily annotated.

At the level of our event hierarchy, these normal-disease pairings support a disease event hierarchy that parallels our normal event hierarchy, a useful and generalizable organization. These pairings also support a visualization scheme that highlights the relationship between the normal and disease processes. As described previously (15), the physical entities and their interactions that



**Figure 1.** Example of disease curation and visualization in Reactome. The normal process of glucose export from the liver under fasting conditions (A) is disrupted by mutations that block glucose-6-phosphate hydrolysis within the endoplasmic reticulum (B) or the transport of glucose-6-phosphate and orthophosphate (Pi) between the endoplasmic reticulum and the cytosol (C).

comprise a pathway are laid out in a pathway diagram that follows the SBGN process description language (<http://www.sbgn.org/Documents/Specifications>) and that is displayed on our Web site. The software that generates these displays has been extended to create disease displays in which the variant forms of the events responsible for a disease process are superimposed and highlighted on the normal process diagram (Figure 1).

We have used these extensions of the Reactome data model and pathway visualization process to create a new disease pathway classification in our event hierarchy that incorporates existing material such as the HIV and influenza life cycles, amyloid formation and botulinus toxin neurotoxicity, together with new material that includes malignant transformation due to mutations in the EGFR and FGFR signaling pathways and mucopolysaccharidoses. Our current release includes annotations for 420 mutant forms of 68 proteins.

### Updated Reactome Web site

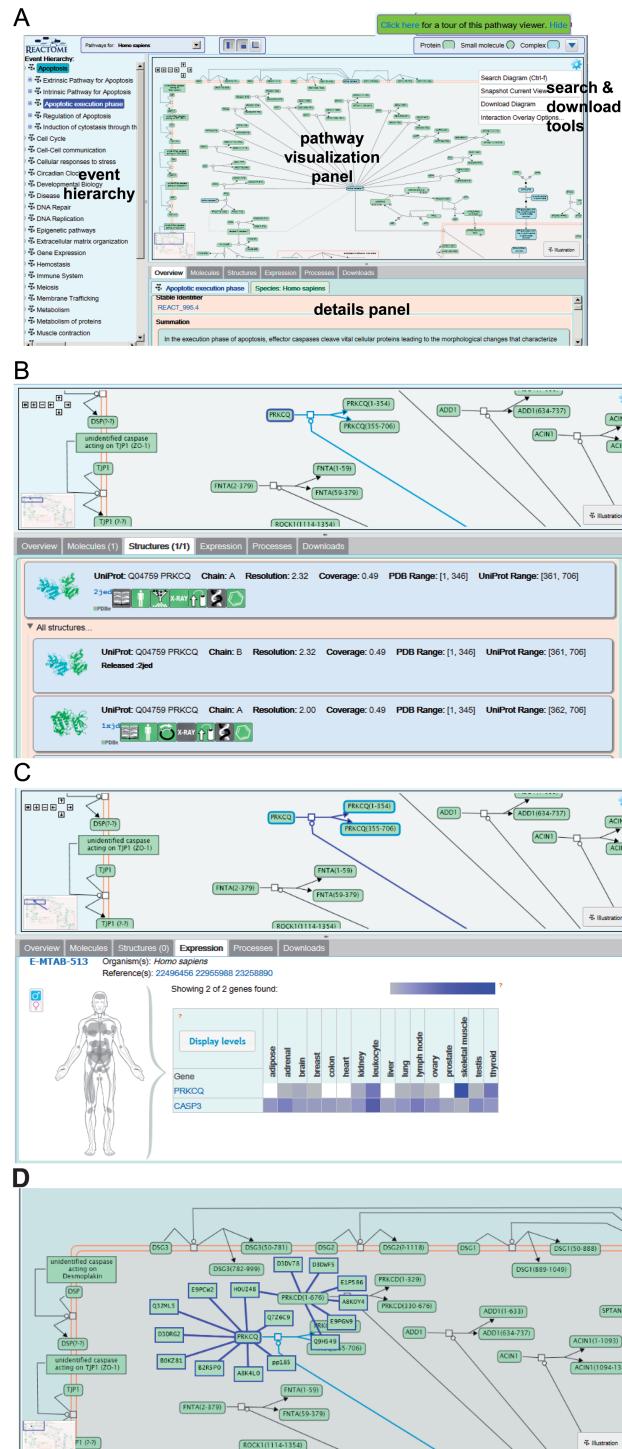
Our home page has been redesigned completely to support intuitive access to our pathway browsing and data analysis tools. The new Web site retains a top menu bar to provide easy access to all of our tools and resources, accompanied by a central panel of links to our most widely used tools and a footer that displays all tools and resources. This organization is consistent with the general model for resources associated with EBI. News is now available as an interactive Twitter display, which also provides open real-time feedback, both by the Reactome group and our user community.

### Dynamic pathway portal

We have improved the flexibility and performance of our pathway browser by creating a new pathway diagram visualization tool using the canvas element introduced in HTML 5. The canvas element is used to render whole pathway diagrams in a Google map-like way with XML-encoded pathway diagram data retrieved from the server using a RESTful API (Figure 2A). The new diagram visualization tool offers quicker performance and better data overlaying technologies (see later in text) and bypasses the slow step to generate static images during database release.

The event hierarchy panel to the left has been redesigned to provide interactivity and access to the entire listing of all the Reactome pathways. Icons now indicate whether a pathway is new (N) or updated (U), and identify ones that are parts of disease processes (+). Navigation controls in the upper left corner enable zooming and panning across the pathway panel. A diagram thumbnail in the lower left shows the part of the pathway currently displayed in the visualization panel. A widget icon in the upper right corner of the panel links to tools for searching within the displayed pathway, overlaying the pathway with functional interactors (see later in text) and for downloading the diagram as a snapshot or PNG file.

A new tabbed ‘Details’ panel, below the pathway diagram, provides additional graphical and textual information. The ‘Overview’ tab provides summary



**Figure 2.** Visualization of the execution phase of apoptosis. (A) Choosing the entry for this event from the event hierarchy in the left panel of the web page causes the pathway to be displayed in a large panel to the right, laid out in SBGN process description format. Buttons at the top of the panel open a diagram key and give access to a brief tour of features of the web page. The panel at the bottom of the page contains text descriptions of various features of the pathway and participating molecules, linked to external databases. (B) The user has selected the reaction ‘Caspase 3-mediated cleavage of PKC delta’; the ‘Structures’ detail tab displays PDB 3D structural data and citations for proteins in that reaction and the ‘Expression’ detail tab (C) displays condition-specific gene expression data from the Gene Expression Atlas. (D) The molecular interaction (MI) overlay displays proteins that interact with user-specified proteins PRKCQ and PRKCD.

information relating to the pathway, reaction or entity selected. The ‘Molecule’ tab displays external annotation and linking-out to other bioinformatic resources describing the selected molecule. The ‘Structures’ tab (Figure 2B) displays 3D structural data and citations for proteins from PDB (16), for small molecules from ChEBI and for the stoichiometry of metabolic reactions from Rhea. The ‘Expression’ tab (Figure 2C) serves condition-specific gene expression data from the Gene Expression Atlas (17). The ‘Processes’ tab displays Reactome pathways and reactions associated with the selected pathway node or event. Finally, the ‘Download’ tab allows users to download the description of the pathway in a variety of formats compatible with third-party tools.

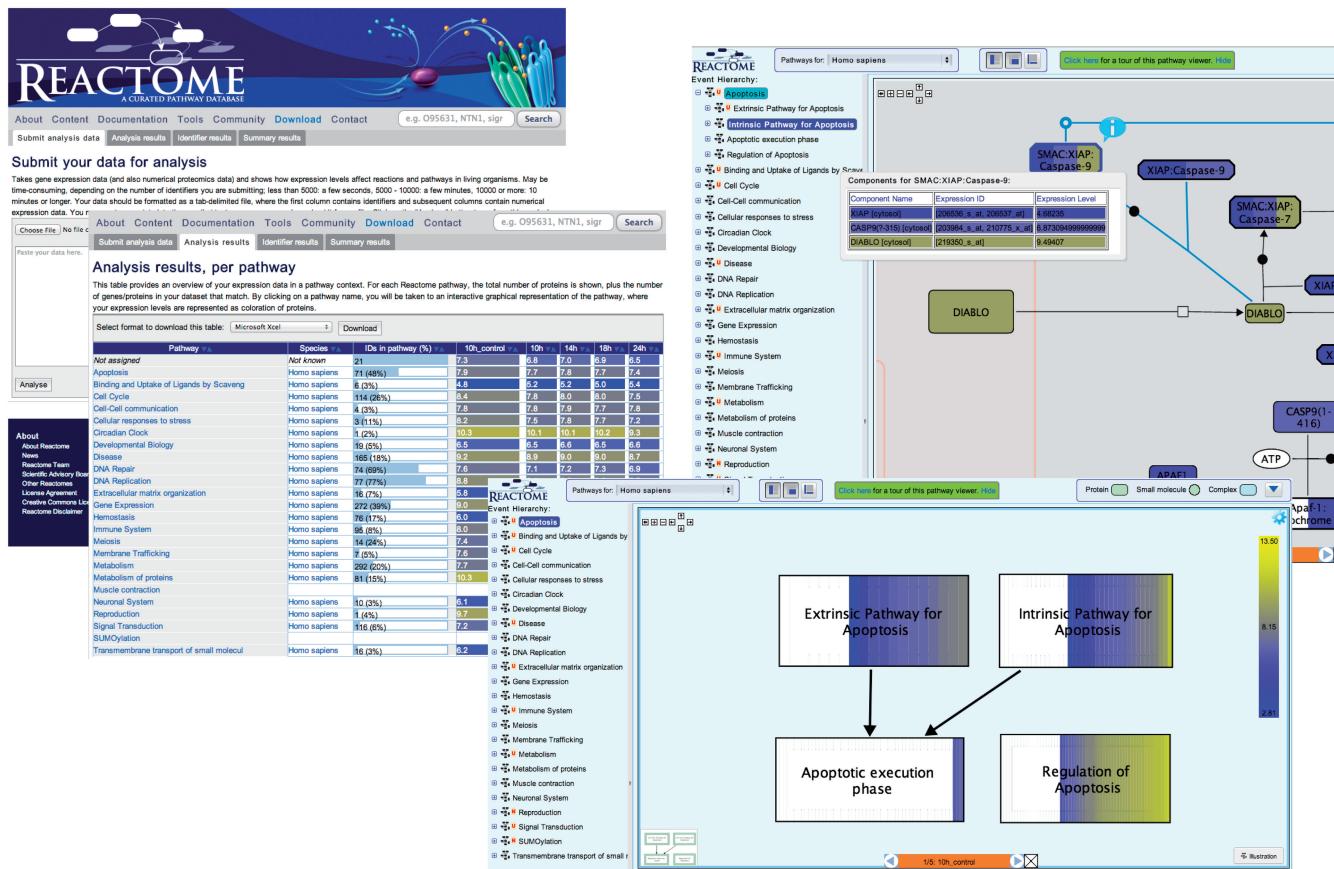
Based on usability testing and feedback from our user community, we modified the context sensitive menus. In this new version, the mouse cursor can be positioned on the pathway background and right-clicked to display a new set of pathway diagram options.

The molecular interaction overlay that provides the display of proteins or small molecules that interact with Reactome pathway proteins (Figure 2D) is revised to include additional interaction databases through the PSIQUIC registry (18) including BindingDB (19), DrugBank (20) and GeneMANIA (21).

## Unified pathway analysis portal

We have merged pathway identifier mapping, over-representation and expression analysis tools into a single tabbed data analysis portal with integrated visualization and summary features (Figure 3). The pathway data analysis module accepts gene lists or expression data with numerical values (e.g. expression, abundance, fold expression change and quality scores). It will automatically distinguish between tab-delimited, comma-delimited or Microsoft Excel files and can cope with files that have been ZIP compressed.

The results of an analysis are displayed in three tabs that provide three different views of the same data. Pathway-oriented expression analysis provides a table of top-level pathways plus information relevant to the type of data submitted by the user. If presented with a list of identifiers, an enrichment analysis will be shown; numerical data will be summarized in a graphical way, using color to indicate average expression values. The identifier mapping view lists each of the identifiers supplied by the user with the corresponding UniProt identifier and a list of pathways in Reactome in which it functions. The overview is a completely new feature. It provides a canvas-based aggregated summary of the entire set of Reactome pathways, using visual cues to inform the user about the



**Figure 3.** Pathway analysis. The data analysis workflow proceeds in three steps: data submission, a tabular display of results and visual display of results as an overlay in a pathway browser window. The fourth pane shows a detail of the pathway browser window, to highlight the display of expression levels.

number of participating entities, relations between the pathways, average expression levels and other details.

The pathway names in the pathway-oriented view are clickable links, which take the user to the relevant pathway diagram. In it, the sub-pathway nodes, complexes and entities are colored according to expression level. Entities that lack gene expression information are colored white. Pathway and complex nodes are colored in vertical segments with each segment representing a protein in the pathway or a component of the complex. The segments are stacked in order from lowest to highest expression level. A new icon at the base of the diagram allows the user to single step through the individual experiments or time points.

### Community relationships and data exchange

Reactome continues to collaborate with other data resources, such as GO, NCBI, EBI and WikiPathways (22). Reactome provides a series of link-outs to many online bioinformatics resources from its protein pages; we have added links to GeneCards annotations (23). Reactome is open-source and open-data, and we have continuously supported the major open-data standards in the domain, including BioPAX levels 2 and 3 (24), PSI-MTAB (18), Protégé (<http://protege.stanford.edu>), SBML-ML (25) and SBGN export format. Reactome now provides an SBGN file format generated using libSBGN for individual pathways. The Reactome SBML export has been upgraded to Level 2, Version 4 and is enriched with a wide variety of additional annotations, including Systems Biology Ontology terms (26). Reactome also supports the Protein Ontology in developing an ontology for protein modifications and protein complexes (27). Our new RESTful API provides outside users with direct access to pathway data in Reactome.

The Reactome data model has been adopted by the Gramene group for manual annotation of plant pathways, especially metabolic processes specific to plants. The current release of Plant Reactome (<http://plants.reactome.org>) includes 131 rice pathways.

Since 2011, Reactome has participated in the Google Summer of Code program as part of the Genome Informatics group, helping to create the software components for the new pathway browser, RESTful API and the pathway overview. The Reactome data and source code continues to be publically accessible under the terms of a Creative Commons Attribution 3.0 Unported License.

### ACKNOWLEDGEMENTS

The authors are grateful to the many expert scientists who have collaborated with us as external authors and reviewers of Reactome content. They thank three anonymous reviewers for their comments on this article.

### FUNDING

National Human Genome Research Institute at the National Institutes of Health [U41 HG003751]; Ontario Research (GL2) fund; European Bioinformatics Institute;

European Commission (PSIMEx); Google Summer of Code Program (2011–2013). Funding for open access charge: National Institutes of Health [U41 HG003751].

*Conflict of interest statement.* None declared.

### REFERENCES

- Biankin,A.V., Waddell,N., Kassahn,K.S., Gingras,M.C., Muthuswamy,L.B., Johns,A.L., Miller,D.K., Wilson,P.J., Patch,A.M., Wu,J. *et al.* (2012) Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. *Nature*, **491**, 399–405.
- Gieger,C., Radhakrishnan,A., Cvejic,A., Tang,W., Porcu,E., Pistis,G., Serbanovic-Canic,J., Elling,U., Goodall,A.H., Labrune,Y. *et al.* (2011) New gene functions in megakaryopoiesis and platelet formation. *Nature*, **480**, 201–208.
- Wu,G., Feng,X. and Stein,L. (2010) A human functional protein interaction network and its application to cancer data analysis. *Genome Biol.*, **11**, R53.
- Vastrik,I., D'Eustachio,P., Schmidt,E., Gopinath,G., Croft,D., de Bono,B., Gillespie,M., Jassal,B., Lewis,S., Matthews,L. *et al.* (2007) Reactome: a knowledge base of biologic pathways and processes. *Genome Biol.*, **8**, R39.
- Kanehisa,M., Goto,S., Sato,Y., Furumichi,M. and Tanabe,M. (2012) KEGG for integration and interpretation of large-scale molecular datasets. *Nucleic Acids Res.*, **40**, D109–D114.
- Karp,P.D., Krummenacker,M., Paley,S. and Wagg,J. (1999) Integrated pathway-genome databases and their role in drug discovery. *Trends Biotechnol.*, **17**, 275–281.
- UniProt Consortium. (2012) Reorganizing the protein space at the universal protein resource (UniProt). *Nucleic Acids Res.*, **40**, D71–D75.
- Flicek,P., Amode,M.R., Barrell,D., Beal,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fairley,S., Fitzgerald,S. *et al.* (2012) Ensembl 2012. *Nucleic Acids Res.*, **40**, D84–D90.
- Hastings,J., de Matos,P., Dekker,A., Ennis,M., Harsha,B., Kale,N., Muthukrishnan,V., Owen,G., Turner,S., Williams,M. *et al.* (2013) The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res.*, **41**, D456–D463.
- Alcántara,R., Axelsen,K.B., Morgat,A., Belda,E., Coudert,E., Bridge,A., Cao,H., de Matos,P., Ennis,M., Turner,S. *et al.* (2012) Rhea—a manually curated resource of biochemical reactions. *Nucleic Acids Res.*, **40**, D754–D760.
- Schriml,L.M., Arze,C., Nadendla,S., Chang,Y.W., Mazaitis,M., Felix,V., Feng,G. and Kibbe,W.A. (2012) Disease ontology: a backbone for disease semantic integration. *Nucleic Acids Res.*, **40**, D940–D946.
- Gene Ontology Consortium. (2010) The gene ontology in 2010: extensions and refinements. *Nucleic Acids Res.*, **38**, D331–D335.
- Milacic,M., Haw,R., Rothfels,K., Wu,G., Croft,D., Hermjakob,H., D'Eustachio,P. and Stein,L. (2012) Annotating cancer variants and anti-cancer therapeutics in reactome. *Cancers*, **4**, 1180–1211.
- Montecchi-Palazzi,L., Beavis,R., Binz,P.A., Chalkley,R.J., Cottrell,J., Creasy,D., Shofstahl,J., Seymour,S.L. and Garavelli,J.S. (2008) The PSI-MOD community standard for representation of protein modification data. *Nat. Biotechnol.*, **26**, 864–866.
- Croft,D., O'Kelly,G., Wu,G., Haw,R., Gillespie,M., Matthews,L., Caudy,M., Garapati,P., Gopinath,G., Jassal,B. *et al.* (2011) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.*, **39**, D691–D697.
- Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.
- Kapushesky,M., Adamusiak,T., Burdett,T., Culhane,A., Farne,A., Filippov,A., Holloway,E., Klebanov,A., Kryvych,N., Kurbatova,N. *et al.* (2012) Gene expression atlas update - a value-added database of microarray and sequencing-based

- functional genomics experiments. *Nucleic Acids Res.*, **40**, D1077–D1081.
18. Aranda,B., Blankenburg,H., Kerrien,S., Brinkman,F.S., Ceol,A., Chautard,E., Dana,J.M., De Las Rivas,J., Dumousseau,M., Galeota,E. *et al.* (2011) PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nat. Methods*, **8**, 528–529.
  19. Liu,T., Lin,Y., Wen,X., Jorissen,R.N. and Gilson,M.K. (2007) BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.*, **35**, D198–D201.
  20. Knox,C., Law,V., Jewison,T., Liu,P., Ly,S., Frolkis,A., Pon,A., Banco,K., Mak,C., Neveu,V. *et al.* (2011) DrugBank 3.0: a comprehensive resource for ‘omics’ research on drugs. *Nucleic Acids Res.*, **39**, D1035–D1041.
  21. Warde-Farley,D., Donaldson,S.L., Comes,O., Zuberi,K., Badrawi,R., Chao,P., Franz,M., Grouios,C., Kazi,F., Lopes,C.T. *et al.* (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.*, **38**, W214–W220.
  22. Kelder,T., van Iersel,M.P., Hanspers,K., Kutmon,M., Conklin,B.R., Evelo,C.T. and Pico,A.R. (2012)
  - WikiPathways: building research communities on biological pathways. *Nucleic Acids Res.*, **40**, D1301–D1307.
  23. Safran,M., Dalah,I., Alexander,J., Rosen,N., Iny Stein,T., Shmoish,M., Nativ,N., Bahir,I., Doniger,T., Krug,H. *et al.* (2010) GeneCards Version 3: the human gene integrator. *Database*, **2010**, baq020.
  24. Demir,E., Cary,M.P., Paley,S., Fukuda,K., Lemer,C., Vastrik,I., Wu,G., D'Eustachio,P., Schaefer,C., Luciano,J. *et al.* (2010) The BioPAX community standard for pathway data sharing. *Nat. Biotechnol.*, **28**, 935–942.
  25. Bornstein,B.J., Keating,S.M., Jouraku,A. and Hucka,M. (2008) LibSBML: an API Library for SBML. *Bioinformatics*, **24**, 880–881.
  26. Courtot,M., Juty,N., Knüpfer,C., Waltemath,D., Zhukova,A., Dräger,A., Dumontier,M., Finney,A., Golebiewski,M., Hastings,J. *et al.* (2011) Model storage, exchange and integration. *Mol. Syst. Biol.*, **7**, 543.
  27. Natale,D.A., Arighi,C.N., Barker,W.C., Blake,J.A., Bult,C.J., Caudy,M., Drabkin,H.J., D'Eustachio,P., Evsikov,A.V., Huang,H. *et al.* (2011) The protein ontology: a structured representation of protein forms and complexes. *Nucleic Acids Res.*, **39**, D539–D545.